FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 10 No. 6, March, 2026, pp 368 – 376
DOI: https://doi.org/10.33003/fjs-2026-1006-4633

EXPLAINABLE MACHINE LEARNING FOR STUDENT ACADEMIC PERFORMANCE PREDICTION IN DATA-CONSTRAINED EDUCATIONAL SETTINGS

***[1]Godwin A. Otu, [2] Federick I. Okonkwo, [3]Lucky I. Okonkwo, [4]Opeyemi O. Adekogba, and [5]Joshua Y. Anche**

[1]Department of Computer Science, Nigerian Defence Academy, Kaduna, Nigeria.
[2]Department of Information and communication Technology, Air Force Institute of Technology Kaduna, Nigeria.
[3]Department of Software Engineering, Veritas University, Abuja, Nigeria.
[4]Department of Telecomunication Engineering, Air Force Institute of Technology Kaduna, Nigeria.
[5]Department of Statistics, Air Force Institute of Technology Kaduna State, Nigeria.

*Corresponding authors' email: gaotu@nda.edu.ng

ABSTRACT

The research explores the effect of a limited dataset in terms of size on machine learning models used for predicting student academic performance in schools. Model generalization can easily be achieved when the dataset size is large. This is not the case in educational data mining, due to the limited data in the field, making model generalization, interpretability, and fairness difficult to achieve. A stratified subsampling technique was used at sizes of 50, 100, 150, 200, and 300 to handle the data scarcity problem. In the same vein, an oversampling strategy was used to balance classes that were not well represented to eliminate bias. SHAP and permutation importance were used to perform interpretability of results, while Spearman rank correlation and Jaccard similarity were used for explanation stability. A fairness audit was also carried out to identify how other socio-demographic factors, other than gender, affect academic performance. The dataset used in this research is the mathematics scores of Portuguese students. Standardized scaling, ordinal encoding, one-hot encoding, and target attribute definition (pass or fail) are performed on the dataset. Logistic Regression, Random Forest, and XGBoost models were trained and evaluated on the dataset. The results from the models were evaluated using accuracy, F1-score, AUC, and Brier score metrics. Results show that Random Forest and XGBoost performed better in terms of accuracy, robustness, and calibration, even with small datasets, when compared to the Logistic Regression model.

Keywords: Student Performance Prediction, Machine Learning, Fairness Auditing, Interpretability Stability, Dataset Constraint

INTRODUCTION

Predicting student academic performance has become very popular, because schools use data-driven methods to identify students that are performing well and also proffer timely solutions (Ngulube, 2025). With artificial intelligence (AI) and the availability of data, educational institutions now use AI in enhancing their decision-making processes and the improvement of student learning outcomes (Esomonu, 2025). Machine Learning (ML), a subsidiary of AI, has been used by researchers to predict student academic outcomes, yet many of these methods have buttressed on the need for predictive accuracy without paying adequate attention to result interpretability, fairness analysis, and the unavailability of large volumes of data in the educational domain (Kesgin *et al*., 2025). The challenges from Kesgin *et al*. (2025) provide a benchmark by combining fairness analysis and explainability into the predictive modeling of student results using the UCI Student Performance dataset.

The research will solve the following challenges: (1) while the benchmark paper used the entire dataset to build the machine learning model, in this research, building and performance evaluation of the model is done with varying input size. (2) Existing research used feature importance analysis for model interpretation; this research will employ SHAP, Spearman rank correlation and Jaccard similarity will be used for explanation stability. (3) Fairness analysis in the benchmark paper was restricted to gender; this research extends fairness audits to many demographic attributes such as parental education attainment and family size. By addressing these challenges, this research will build a framework that will predict student academic performance with limited data.

Literature Review

The use of different machine learning models, datasets, and interpretability techniques has shaped the field of education data mining. A review of related work in the area of predictive accuracy, fairness, and explainability in educational machine learning, will provide a research direction.

Machine Learning Approaches for Predicting Student Success in Education

Machine learning models are powerful tools used for the analysis of different data derived from different sources. These models perform based on the quality of data given to them. Data used in educational data mining can be categorized into academic records, demographic features, behavioural indicators, and socio-economic backgrounds. Some machine learning models used in predictive analytics are logistic regression, random forest, support vector machine, gradient boosting and deep learning models.

Logistic Regression*:* Logistic Regression is one of the most widely used models in educational prediction problems because of its reliability when it comes to result interpretability and model efficiency. It offers direct insights into the influence of individual attributes, such as study time or parental education, on the likelihood of student success. The model approximates the probability of a student passing or failing as a logistic function of the input attributes, expressed as:

$$P(Y = 1 \mid X) = \frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n)}} \quad (1)$$

From the equation, P(Y=1|X) is the probability of success, β_0 is the intercept, and β_i are the coefficients which represent the contribution of each attribute X_i (Taylanova,

2024). The simplicity of Logistic Regression is preferred in research where transparency and explainability are important for policymakers in education.

Random Forests: Random Forests model show great capability because they used ensemble learning technique. This technique integrates many decision trees to improve the accuracy and robustness of the model. Random Forests are effective in the area of educational data mining due to their being able to capture intricate and non-linear relationships between attributes. This makes random forest models suitable for classifying heterogeneous datasets. The prediction by random forest classifier is achieved by aggregating the outputs of individual decision trees using majority voting, which is expressed as:

$$\hat{y} = mode\{h_1(X), h_2(X), \ldots . h_T(X)\} \quad (2)$$

Where $h_t(X)$ denotes the prediction of the t^{th}decision tree, and T is the total number of trees in the forest (Chandralekha *et al.* 2025). Random Forests also offer attribute importance measures by averaging the reduction in impurity across trees, thereby aiding in the identification of the most influential predictors that assist in student academic performance prediction tasks.

Support Vector Machines (SVM): Support Vector Machines (SVMs) have been applied in the prediction of student academic success, mostly when the datasets are high-dimensional. The widening of the margin between classes, SVM models provide strong performance in distinguishing between successful and unsuccessful students. The decision function of a linear SVM can be written as:

$$f(x) = sign(\sum_i^N \alpha_i\, y_i K(x_i, x) + b \quad (3)$$

Where $x_i$are the support vectors, $y_i \in \{-1, +1\}$ are target labels, $\alpha_i$are Lagrange multipliers, $K(x_i, x)$ is the kernel function, and b is the bias term (Zollanvari, 2023). Widening the margin between classes, makes the SVMs attain a high generalization performance, but not without an interpretability limitation when it is compared to simpler models such as Logistic Regression.

Dient Boosting Techniques*:* The use of Gradient Boosting techniques which include XGBoost and LightGBM, have become very popular due to their exceptional predictive accuracy. These models sequentially improve weak learners by focusing on difficult-to-predict instances, making them highly capable in handling educational datasets with intricate patterns. The general gradient boosting model builds an additive function by repetitively fitting weak learners to the negative gradients (residuals) of a loss function, given as:

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \quad (4)$$

Where $F_m(x)$ is the ensemble model at iteration m, $h_m(x)$ is the weak learner (typically a decision tree), and $\gamma_m$is the learning rate scaling factor. The weak learner is fitted to the pseudo-residuals, expressed as:

$$\gamma_{im} = -[\frac{\delta L(y_i, F(x_i))}{\delta F(x_i)}]_{F(x)=F_{m-1}(x)} \quad (5)$$

With L representing the loss function (Zollanvari, 2023). Gradient Boosting attains fine-grained optimization with the aid of loop refinement. Hyperparameters (learning rate, maximum depth, and regularization terms) in gradient boosting control model intricacy and enhance generalization, but not without additional computational demands.

Deep Learning (DL)*:* DL model is the most advanced class of models used in educational data mining. The model has a complex network architecture called neural network. These layers of complex design can capture complex attribute interactions and temporal patterns in educational dataset. Although deep learning models most times achieve superior accuracy, but their black-box nature gives room for concerns about result interpretability, which is important in educational decision-making. Nonetheless, with new advances in explainable artificial intelligence this challenge is beginning to be a thing of the past, making deeper models to be adopted more responsibly.

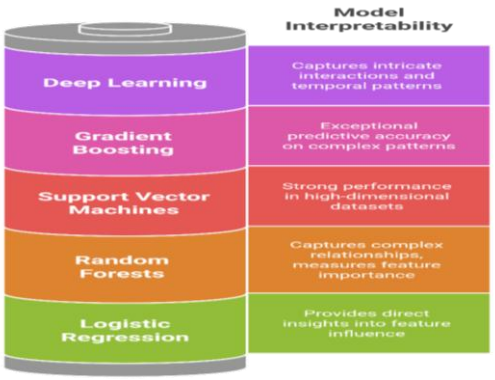

Figure 1: Machine Learning Models Balance Accuracy and Interpretability in Education

Explainability and Interpretability in Educational AI

The adoption of artificial intelligence for educational data mining is because of the availability of frameworks that can readily interpret output from models. This allows policymakers to trust and act upon model predictions. Transparent models assist to ensure fairness, reduce bias, and promote responsible use of AI in decision-making. Many approaches are utilized to offer interpretability in this research.

SHapley Additive exPlanations known as SHAP for short is a framework employed to assign contribution values to individual attributes. SHAP provides not only consistent but a theoretical basis into how features affect predictions. The SHAP value for an attribute i is expressed as:

$$\emptyset_i = \sum_{S \subseteq F(i)} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f(S \cup \{i\}) - f(S)] \quad (6)$$

Where F is the full set of attributes, S is a subset excluding attribute i and f(s) denotes the model prediction using subset S. This formula ensures a fair prediction on how each attribute contributes to student academic success (Wang and Tris, 2025).

LIME (Local Interpretable Model-Agnostic Explanations) offers a localized estimations of model behaviour, pointing out which attributes influence individual predictions, though with less stability across samples. Permutation importance is also employed as a model-agnostic technique, quantifying feature relevance by measuring performance drops when attribute values are randomly shuffled. The permutation importance of a feature j can be written as:

$$PI_j = \frac{1}{R}\sum_{r=1}^{R} M(D) - M(D_{\pi_j}(r)) \quad (7)$$

Where $M(D)$ the performance of the model on the dataset is, $D_{\pi_j^{(r)}}$ is the dataset with feature j permuted in repetition r, and R is the number of permutations (Biswas *et al*., 2025).

For linear models such as Logistic Regression, model coefficients directly provide interpretable weights that reflect the direction and magnitude of each feature's impact.

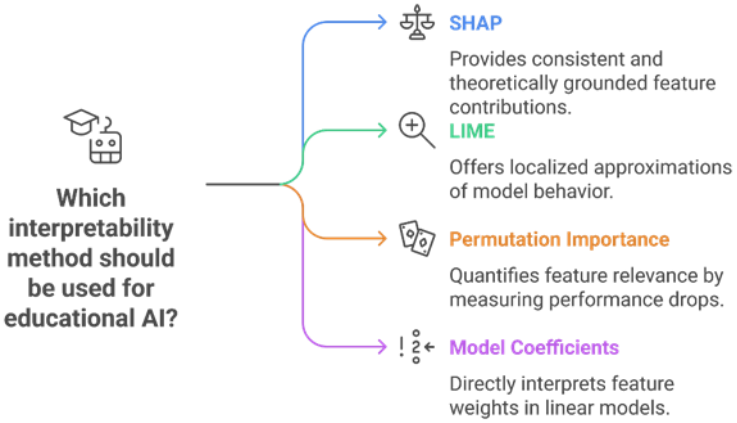

Figure 2: Explainability and Interpretability in Educational AI

Review of Related Work

The use of machine learning to predict academic performance has become very significant in the advancement of students' result outcome by suggesting how students study and improvement of retention capability. Many frameworks and models have been developed in the field of educational data mining, but the scarcity of data has always been a challenge in developing dependable and reliable models (Lünich, and Keller, 2024). The research Investigate performance analysis in universities using ten regression models across two datasets that change in size and feature complexity. Their ensemble voting regression model, combining with the top five models, performs better than standalone models with superior accuracy and minimal error rates. It was seen from the result that the combination of LIME and SHAP improved interpretability. Also, looking beyond accuracy, ethical consideration in educational data mining was buttressed. The research looked at transparency and fairness in educational analytics with the aid of a factorial survey experiment that involved German students. The result opined that model simplicity positively influences perceptions of fairness while accuracy plays a minimal role. The research shows the significance of explainability in predictive analytics.

Ramaswami *et al.* (2022) show how CatBoost when combined with other interpretable methods can provide recommendations for weak students. The research illustrates how descriptive prediction can offer proactive intervention for students. Kalita *et al.* (2025) use Bi-LSTM a deep learning model to predict GPA for students in their second semester. They used CatBoost and XGBoost and achieved an accuracy of 88.23%, incorporated SHAP for interpretability.

To address the challenges associated with fairness, Kesgin *et al*. (2025) apply the oversampling technique and cross-validation on the UCI student dataset with Logistic Regression, Random Forest, and XGBoost models. XGBoost achieved the best results. Raftopoulos *et al.* (2024) handle fairness in admission portals with the aid of disparate impact metrics; this ensures transparent decision-making. Sanfo, (2025) utilizes supervised models on student data from Burkina Faso. The research shows that KNN and SVM are effective in classification tasks, while Random Forest performs better in regression tasks. SHAP analysis in the research showed that community involvement and infrastructure as critical determinants of students' academic performance. In all, it can be seen that while predictive accuracy remains indispensable, interpretability and fairness are very important in educational machine learning studies.

MATERIALS AND METHODS

The model framework begins with the combination of the Mathematics and Portuguese student achievement datasets. The dataset is preprocessed to define the predictive class which take the form of either pass (final grade G3 ≥ 10) or fail (G3 < 10). Feature preprocessing is carried out with a column transformer that performs three transformations at the same time. All numeric variables, which include age, first and second period grades (G1, G2), number of past failures, and school absences, are standardized using the standard scalar library with a default mean and variance normalization. Ordinal attributes, including study time (1–4) and travel time (1–4), are encoded using an ordinal encoder with integer mappings depicting their natural order. All nominal categorical features, such as school, sex, family size, parental education, parental occupation, guardian, and support indicators, are processed with OneHotEncoder; this ensures that unseen categories during testing are handled without error. The resulting preprocessing and modeling pipeline is pictured in Figure 3.

The strategy used to handle data-constrained challenge is subsampling the training data at sizes of N = 50, 100, 150, 200, 300, with each sample size repeated across five random seeds (0–4). A stratified split maintains an 80/20 partition between training and test sets, where the test is not dynamic across all experiments. In cases where minority classes in subsampled sets contain fewer than five samples, a simple oversampling approach is applied by replicating the minority class to achieve parity with the majority class. This step

ensures robustness of training in extremely small-data conditions.

The Logistic Regression, Random Forest classifier and XGBoost classifier models are trained for this research. The performance of each model is evaluated on the fixed test set using accuracy, F1-score, AUC, and the Brier score metrics.

Fairness auditing is integrated into the evaluation pipeline using two metrics, namely; (1) the demographic parity gap (DP), calculated as the absolute difference in positive rates predicted between groups. (2) The equalized odds difference (EO) calculated as the sum of absolute differences in true positive rates and false positive rates across groups. Protected attributes evaluated include gender (male vs female), parental education (low education ≤ 2 vs higher), and family size (≤ 3 vs > 3). An additional intersectional fairness analysis considers the compound attribute of male students with low parental education. These fairness metrics are computed for each trained model across all experimental runs.

The model interpretability is implemented using SHAP (SHapley Additive exPlanations), where tree explainer is applied to random forest and XGBoost models, and linear explainer is applied to logistic regression models. Any time SHAP fails due to sampling limitations, permutation importance with 20 repeats is used as an alternative. To control runtime in small-data settings, kernel explainer is used only for subsets of 50 background and test samples, with a limit of 50 sampling repetitions. Feature importance vectors are obtained from these explanations and are saved per run. Stability of interpretability is quantified by computing the median Spearman rank correlation across feature importance vectors from different seeds and by calculating the Jaccard similarity index of the top five features across runs. Both stability metrics are aggregated across N to demonstrate how explanation consistency improves with larger sample sizes.

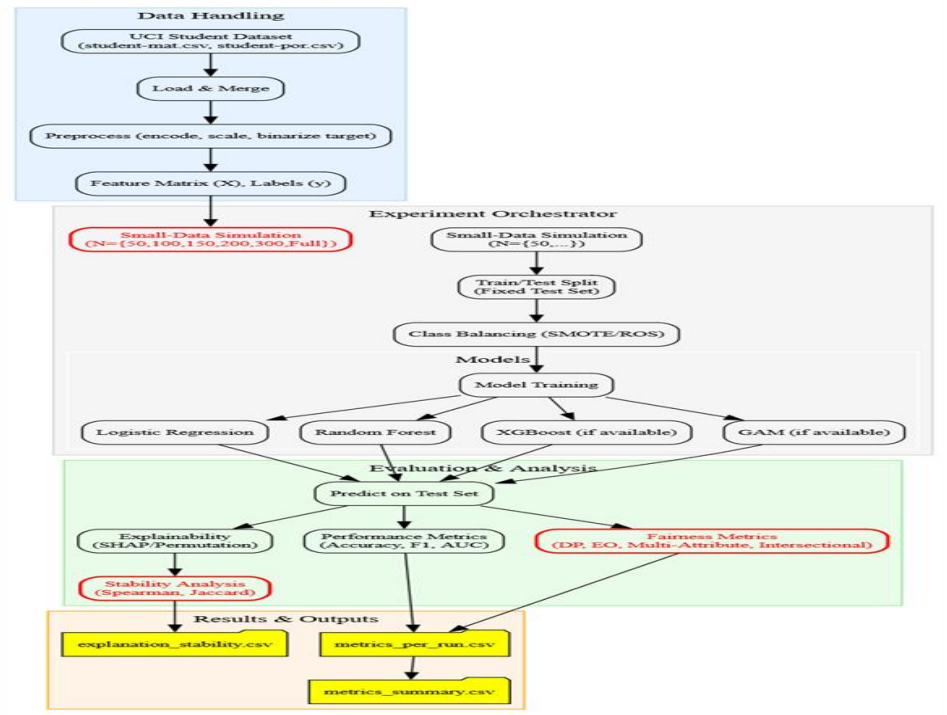

Figure 3: Model Architectural Flow

Figure 3 is the architectural workflow of the research. The pipeline starts with the loading of dataset, preprocessing, and feature-labeling. The experimental framework has three blocks, namely; (1) the simulation of small dataset size to reflect limited student records in developing regions (2) stability analysis of feature explanations in order to ensure consistency across runs and (3) fairness audits expanded to multiple demographic and intersectional attributes. The Models are trained and evaluated under varying sample sizes, with results aggregated into performance, fairness, and stability reports for comparative analysis

The experimental design explicitly subsampled training sets with the repeated random seeds. The results in the tables and the plots clearly show how accuracy, F1, AUC, and Brier vary as training size increases. This addresses the first gap by showing how model reliability changes under data-constrained scenarios. Table 1 and Table 2 present the metrics summary and metrics_per_run, respectively, while figure 4, 5, 6 and 7 visualize accuracy_mean_vs_N, f1_mean_vs_N, auc_mean_vs_N, brier_mean_vs_N.png respectively. These visualizations show how performance improves with larger training sizes, demonstrating reliability under constrained data.

RESULTS AND DISCUSSION

Table 1: Metrics Summary

Model	N	Accuracy	F1-Score	AUC	Brier
Logistic	50	0.915	0.947	0.972	0.057
Logistic	100	0.933	0.958	0.976	0.048
Logistic	150	0.942	0.963	0.982	0.042

Model	N	Accuracy	F1-Score	AUC	Brier
Logistic	200	0.950	0.969	0.984	0.038
Logistic	300	0.957	0.973	0.988	0.032
Random Forest	50	0.972	0.983	0.997	0.037
Random Forest	100	0.991	0.995	1.000	0.023
Random Forest	150	0.994	0.996	1.000	0.020
Random Forest	200	0.994	0.996	1.000	0.016
Random Forest	300	0.997	0.998	1.000	0.012
XGBoost	50	1.000	1.000	1.000	0.002
XGBoost	100	1.000	1.000	1.000	0.001
XGBoost	150	1.000	1.000	1.000	0.000
XGBoost	200	1.000	1.000	1.000	0.000
XGBoost	300	1.000	1.000	1.000	0.000

Across all the training dataset sizes, the XGBoost model performed better than the Logistic regression and the Random forest models. The model achieved a perfect score of 1.00 for accuracy, F1, and AUC with near-zero Brier scores of 0.002 when the dataset size is 50. The Brier score then decreased to 0.000 as the size of the dataset increased to 300. This indicates good predictive accuracy and calibration. This performance, however, may suggest overfitting in small-data scenarios, when the standard deviation is zero, reflecting no variability across runs. Random Forest also exhibited strong performance, with accuracy improving from 0.972 at a dataset size of 50 to 0.997 as the size of the data moved to 300, F1-score moved from 0.983 to 0.998, and AUC from 0.997 to 1.000. The Brier score for the random forest model decreased from 0.037 to 0.012, which shows an improvement in the calibration with larger datasets. The standard deviations for the Random Forest model range from 0.002 to 0.015, while the accuracy and F1 results indicate stable performance with slight variability at smaller dataset sizes. Logistic Regression model, though less accurate than the ensemble models, showed steady improvement, with accuracy rising from 0.915 when the dataset size is 50 to 0.957 when the size of the data is 300, F1-score from 0.947 to 0.973, and AUC from 0.972 to 0.988. The Brier score of the logistic regression model improved from 0.057 to 0.032; this means there was a better calibration as data size increased, though with slightly higher standard deviations of 0.021 for accuracy.

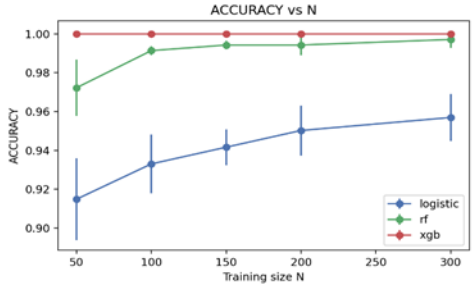

Figure 4: Accuracy Mean vs N

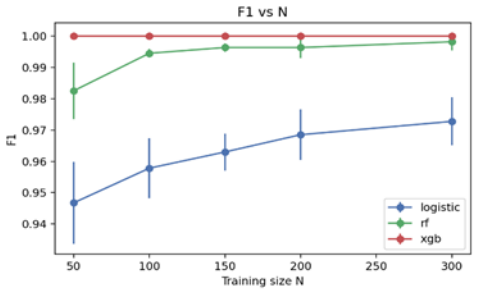

Figure 5: F1 Score vs N

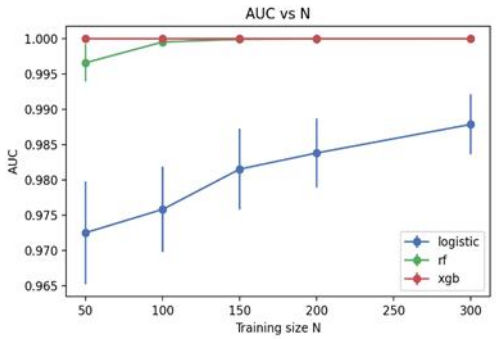

Figure 6: AUC vs N

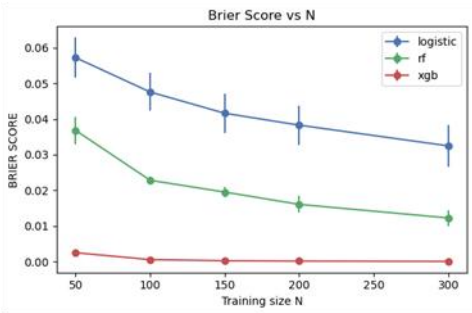

Figure 7: Brier Score vs N

Figures 4 to 7 show the results from the experiment conducted with a moderate amount of data. Starting from Figure 4, the mean accuracy for the Logistic Regression model produced a steady upward trend, increasing approximately from 0.915 when the dataset size was 50 to around 0.957 when the size of the data increased to 300. This shows that the model is very sensitive when the size of the dataset increases. Random Forest model achieves a very high initial accuracy of 0.972 even with a small dataset size of 50 and stabilizes near 0.997 when the size of the data increases to 300. Similarly, the XGBoost model achieves a perfect accuracy of 1.0 across all training sizes. This implies that the model shows a strong learning capacity even in limited data settings.

A similar pattern for the F1-score is observed in Figure 5, where the Logistic Regression model improves gradually from 0.947 to 0.93. The Random Forest model produced

better F1-score values exceeding 0.98 from the smallest training size and approaching 0.998 with large training sizes. XGBoost model once again achieves consistently perfect classification across all sizes.

Figure 6 presents the Area under the ROC Curve (AUC). Logistic Regression model reveals steady gains, rising from 0.972 at a dataset size of 50 to 0.988 when the dataset size became 300, whereas the Random Forest model quickly approaches near-perfect discrimination from 0.996 to 1.0. XGBoost maintains an AUC of 1.0, which cuts across all training sizes, reinforcing its capacity to achieve flawless separation even with minimal data.

The calibration results are represented in Figure 7 using the Brier score, where lower values indicate better probabilistic predictions. Logistic Regression decreases from 0.057 at N=50 to 0.032 at N=300, pointing out improved calibration as training size grows. Random Forest achieves markedly lower Brier scores, improving from 0.037 to 0.012, while XGBoost consistently maintains near-zero error of 0.002 when the dataset size is 50, converging to a value lesser than 0.001 at larger dataset sizes.

The results show how model reliability evolves under constrained data conditions. Logistic Regression model requires larger samples to achieve stability, while Random Forest shows strong performance even with smaller subsets. XGBoost model achieved near-perfect predictive outcomes across all conditions; this suggests that ensemble boosting methods are suited for data-limited educational settings.

Enhancing Interpretability Stability

The research addressed enhanced interpretability with the inclusion of stability analysis. Table 3 reports the explanation stability results, where the Spearman Median captures the rank-order consistency of attribute importance vectors, while the Jaccard Median shows the overlap of the top-five most important attributes across runs. These metrics offer quantitative evidence of how consistent explanation techniques remain under repetitive sampling. Figures 8 and 9 illustrate the Spearman Median and Jaccard Median plot versus training dataset size. These results show if the explanations delivered to educators are robust and can be reproduced in a data-constrained scenario.

Table 3: Explanation Stability

Model	N	Spearman Median	Spearman IQR	Jaccard Median	Jaccard IQR
logistic	50	0.449	0.237	0.429	0.000
logistic	100	0.535	0.075	0.429	0.179
logistic	150	0.496	0.139	0.548	0.238
logistic	200	0.413	0.223	0.548	0.238
logistic	300	0.475	0.339	0.667	0.238
rf	50	0.516	0.111	0.429	0.000
rf	100	0.575	0.073	0.429	0.238
rf	150	0.642	0.102	0.667	0.000
rf	200	0.606	0.122	0.667	0.000
rf	300	0.660	0.064	1.000	0.333
xgb	50	1.000	0.000	1.000	0.000
xgb	100	0.500	0.261	0.250	0.273
xgb	150	1.000	0.000	1.000	0.000
xgb	200	0.718	0.479	0.667	0.806
xgb	300	1.000	0.000	1.000	0.000

The stability analysis using interpretability techniques offers model explanations under varying training sizes. For the logistic regression model exhibits fair consistency, with Spearman Median values ranging from 0.413 to 0.535. Although the rank-order agreement of attribute importance across iterations dipped, the Jaccard Median showed an upward trend, starting from 0.429 when the training sizes are small to 0.667 when the dataset size becomes 300. As illustrated in Figure 8, the Spearman Median values reveal that stability improves incrementally. Figure 9 shows that the Jaccard Median becomes more robust when the dataset size increases. This suggests that the logistic regression model explanations' stability is a function of increased dataset size.

In contrast, the random forest model showed a more robust and consistent stability result. Spearman Median values increased steadily from 0.516 at dataset size of 50 to 0.660 as the size increased to 3, while the Jaccard Median improved from 0.429 to a perfect overlap of 1.000. The narrowing interquartile ranges further depict that random forest model explanations not only became more reproducible with increasing training size but also displayed lower variability across repetitions. Figures 8 and 9 jointly point out the superior stability of random forest model explanations when compared to the logistic regression model. XGBoost presented a distinctive pattern, alternating between perfect stability and instability. At training sizes of 50, , and , both the Spearman and Jaccard Medians achieved perfect values of 1.000; this indicates a complete reproducibility of feature rankings and overlaps across iterations.

However, at dataset size of 100 and 200 shows a dramatic decrease in stability, with Spearman values of 0.500 and 0.718 with Jaccard values of 0.250 and 0.667, followed by wide interquartile ranges. As visualized in Figure 8 (Spearman vs N) and Figure 9 (Jaccard vs N), these fluctuations suggest that XGBoost explanations can be highly stable under certain conditions but may also be sensitive to the specific data partitions when the sample size is modest.

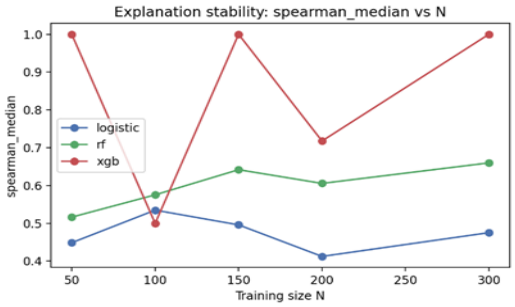

Figure 8: Spearman vs N

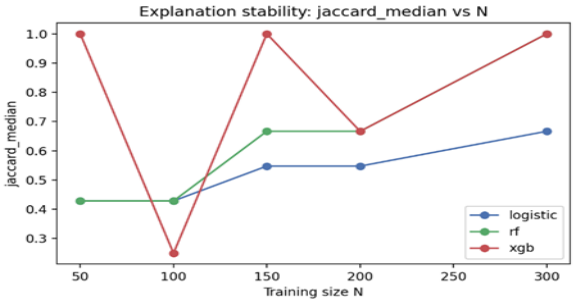

Figure 9: Jaccard vs N

Figure 8 shows a plot of the Spearman Median values versus training size, which illustrates the rank-order consistency of feature importance across iterations. Logistic regression model shows a fluctuating stability that improves slightly as the dataset size increases. Random forest model produced an upward trend pattern, with an increasing stability as training size grows; this indicates more reliable explanations in large dataset size. XGBoost model, however, alternates between perfect stability and reduced stability, reflecting how the model is sensitive to sample size and a partitioned dataset. Figure 9 shows the Jaccard Median values, which capture the intersection among top-ranked attributes across iterations. Logistic regression model again shows average consistency with a marginal improvement as the size of dataset increases while random forest model explanations become evidently stable, achieving near-perfect overlap at higher training sizes. XGBoost model follows the same alternating pattern as in the Spearman results, with values changing between complete stability and noticeable instability depending on the training size.

Broadening Fairness Audits

The fairness evaluation aside genders also captures socio-economic, family-related variables, along with intersectional fairness. Table 1 presents Demographic Parity (DP) and Equalized Odds (EO) values for all subgroups across all iterations, models, and training sizes. The subgroups analyzed include sex, which is male or female, the level of education attained by parents which is labeled low, high and Medium, and the size of the family designated either ≤3 or >3 while the intersection of gender with low parental education. Figures 10 through 17 visualize these metrics, showing DP and EO gaps for each subgroup as training size increases. These plots show how fairness gaps change as training data grows, pointing out if biases are either reduced or increased under small-data conditions. The extension of fairness analysis beyond gender to incorporate socio-economic status, size of family, and intersectional effects, offers a broader and more realistic assessment of inequality in educational predictive analytics.

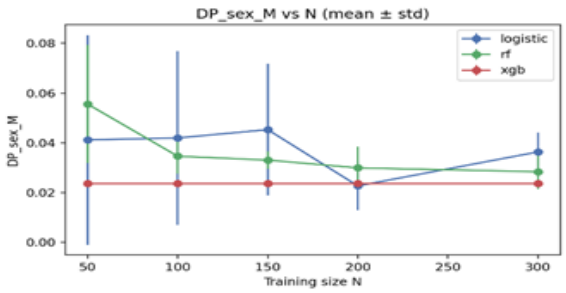

Figure 10: DP_Sex_M vs N

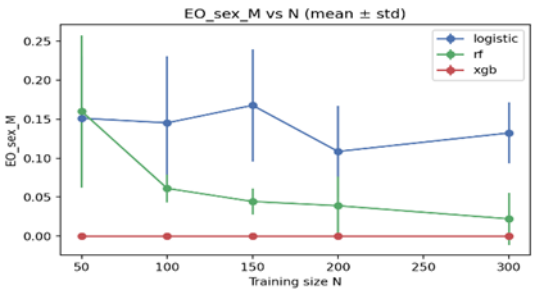

Figure 11: EO_Sex_M vs N

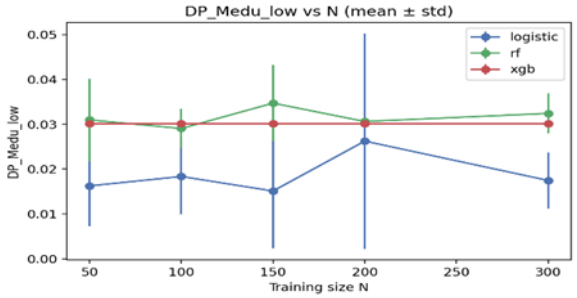

Figure 12: DP_Medu_Low vs N

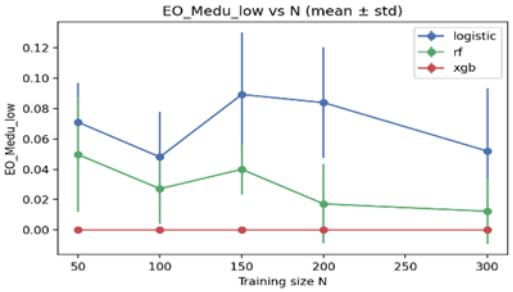

Figure 13: EO_Medu_Low vs N

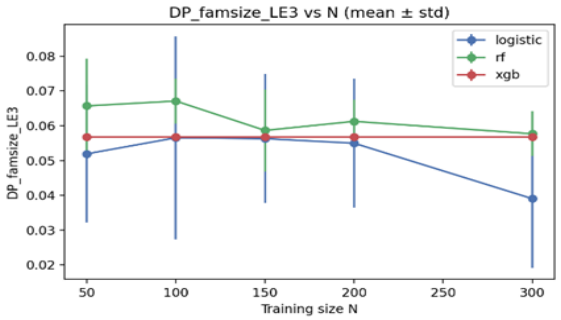

Figure 14: DP_Famsize_LE3 vs N

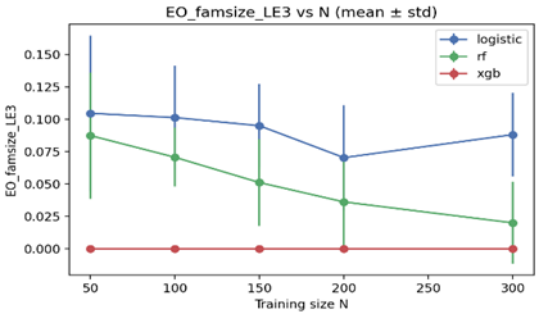

Figure 15: EO_Famsize_LE3 vs N

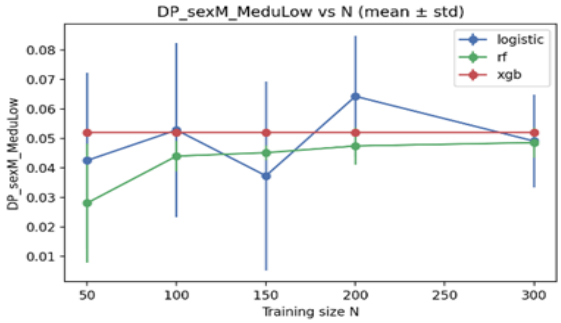

Figure 16: DP_Sex M_Medu Low vs N

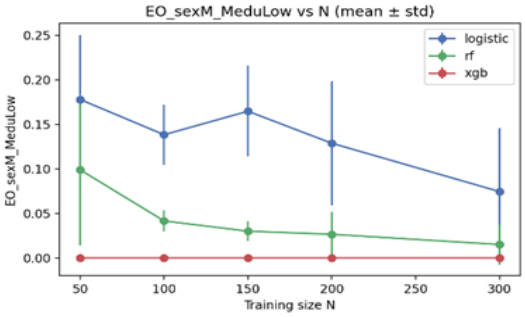

Figure 17: EO_Sex M_Medu Low vs N

The fairness analysis goes beyond gender to incorporate the educational background of parents, the size of the family, and intersectional subgroups. The result for the fairness audit is shown in Figures 10 to 17. The comparison based on gender is shown in Figures 10 and 11, the figures show that demographic parity and equalized odds gaps were higher with small training sets; this can be seen vividly in plots generated by logistic regression model, but these gaps continue to diminish as training size grew, with Random Forest and XGBoost models getting more equitable predictions. The fairness analysis for the level of education by parents is depicted in Figures 12 and 13. The disparities could be clearly seen at low data sizes but gradually declined as the dataset size increases. This asserts that the level of education of parents is an important fairness attribute. The size of the family comparisons is shown in Figures 14 and 15. The fluctuating gaps at smaller training sizes, for both demographic parity and equalized odds differences, decreased with an increase in sample size. The ensemble models performed better than the logistic regression model. The intersectional subgroup of male students whose parents were not degree holders is shown in Figures 16 and 17, with the largest fairness gaps under small-data sizes, affirming that compounded demographic disadvantages increase bias. However, these gaps also reduce to the bare as training sizes grew, reinforcing the importance of sufficient data in mitigating inequities. These results show that while small-data contexts exacerbate fairness risks across gender, socio-economic, and family-based dimensions, ensemble models when trained with larger datasets can substantially reduce bias. When bias is reduced, the fairness gap can be addressed, which in turn offers a broader and more realistic fairness evaluation.

Results

Structured by Objectives

The first research gap is addressed by training dataset using subsample sizes of 50, 100, 150, 200, and 300 students. The output is given in Table 1 and Figures 4 to 7, which depicts how predictive reliability depends on data size. The logistic regression model displayed steady but moderate improvements in accuracy, F1-score, and AUC with larger datasets. Random Forest had exceeding scores across all training sizes. XGBoost achieved an accuracy, F1, AUC score close to 1.0, with very low Brier scores. The downward trends in Brier scores across models show improved calibration as training size grew. These findings assert that model performance is sensitive to training size, with ensemble techniques providing the most reliable outcomes with small sized data. Secondly, Spearman rank correlation was used to address model explanation stability while similarity of top-k attributes across repeated runs was handled by using Jaccard strategy. Table 2 and Figures 8 to 9 show that stability changes across models and training sizes. Logistic regression produced moderate variable stability, with Spearman medians covering from 0.41 to 0.54 and Jaccard overlap improving slightly when the dataset size increases. Random Forest attained higher and a more consistent stability, mostly as the training sizes increased, where Jaccard medians reached 1.0; this indicates a complete overlap in top attributes across runs. XGBoost was getting to a near-perfect stability in many cases, though with fluctuations in training sizes. These results offer direct numerical evidence that explanation stability improves as the training size increases and is model-dependent; this ensures that interpretability of results in educational data mining can be achieved across varying data constraints. The result which shows how fairness evaluation could be evaluated beyond gender to incorporate the educational level of parents, the size of the family, and intersectional subgroups is shown in Table 3 and Figures 10 to 17. The figures show the demographic parity (DP) and equalized odds (EO) gaps across models and training sizes. Gender-based fairness gaps were very large for the logistic regression model under small-data conditions but decreased as training data size grew, with Random Forest and XGBoost offering more equitable predictions. Similar improvements were seen for parental education, where gaps were significant at low N but stabilized with more data. Family size comparisons showed fluctuations

at smaller training sizes, but fairness gaps diminished substantially as sample sizes increased. The intersectional subgroup of male students with low parental education displayed the most pronounced disparities under small-data conditions, asserting the compounding nature of demographic disadvantages. However, as with other subgroups, these gaps declined with larger training sets, showing the mitigating effect of more data. Together, these results offer a broader fairness analysis than the benchmark, making sure evaluations capture the intricate realities of inequality in education.

CONCLUSION

The research addressed three critical gaps in student academic performance prediction. Through the simulation of small-data, the research asserts that machine learning models output is a function of the training size. Random Forest and XGBoost offered superior performance and calibration in constrained data conditions. The addition of interpretability methods in the research proves that consistent and reproducible explanations can be achieved. Moreover, the expanded fairness audits show that predictive disparities spread from gender to socio-economic factors. Collectively, these contributions provide a robust and comprehensive framework for machine learning model deployment to schools where data scarcity, interpretability, and fairness remain a major problem yet to be addressed. Further research can explore the use of big data to evaluate the generalizability of the framework across different educational institutions.

REFERENCES

Ahmed, W., Wani, M. A., Plawiak, P., Meshoul, S., Mahmoud, A., & Hammad, M. (2025). Machinelearning-based academic performance prediction with explainability for enhanced decision-making in educational institutions. *Scientific Reports*, *15*(1), 26879.

Biswas, S., Grundlingh, N., Boardman, J., White, J., & Le, L. (2025). A Target Permutation Test for Statistical Significance of Feature Importance in Differentiable Models. *Electronics*, *14*(3), 571. https://doi.org/10.3390/electronics14030571

Chandralekha, E., Dhineesh, I., Reddy, G. L., & Ganesh, T. (2025, June). IoT-Enabled Device for Predictive Monitoring and Disease Management in Cow. In *2025 3rd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)* (pp. 531-537). IEEE.

Esomonu, N. P. M. (2025). Utilizing AI and Big Data for Predictive Insights on Institutional Performance and Student Success: A Data-Driven Approach to Quality Assurance. *AI and Ethics, Academic Integrity and the Future of Quality Assurance in Higher Education*, *29*.

Kalita, E., Alfarwan, A. M., El Aouifi, H., Kukkar, A., Hussain, S., Ali, T., & Gaftandzhieva, S. (2025, June). Predicting student academic performance using Bi-LSTM: a deep learning framework with SHAP-based interpretability and statistical validation. In *Frontiers in Education* (Vol. 10, p. 1581247). Frontiers Media SA.

Kesgin, K., Kiraz, S., Kosunalp, S., & Stoycheva, B. (2025). Beyond Performance: Explaining and Ensuring Fairness in Student Academic Performance Prediction with Machine Learning. *Applied Sciences*, *15*(15), 8409.

Lünich, M., & Keller, B. (2024, June). Explainable artificial intelligence for academic performance prediction. An experimental study on the impact of accuracy and simplicity of decision trees on causability and fairness perceptions. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1031-1042).

Ngulube, P. (2025). Predicting Academic Success and Identifying At-Risk Students: A Systematic Review of Data Analytics and Machine Learning Approaches in Higher Education Institutions. *Educational Administration: Theory and Practice*, *31*(1), 117-134.

Raftopoulos, G., Davrazos, G., & Kotsiantis, S. (2024). Fair and transparent student admission prediction using machine learning models. *Algorithms*, *17*(12), 572.

Ramaswami, G., Susnjak, T., & Mathrani, A. (2022). Supporting students' academic performance using explainable machine learning with automated prescriptive analytics. *Big Data and Cognitive Computing*, *6*(4), 105.

Sanfo, J. B. M. (2025). Application of explainable artificial intelligence approach to predict student learning outcomes. *Journal of Computational Social Science*, *8*(1), 9.

Taylanova, S. Z. (2024). The Rationale for the Present Study is based on the Following Pedagogical Conditions for Developing Students' Technical Thinking in English Language Classes. Best Journal of Innovation in Science, Research and Development, 3(12), 101-110.

Wang, X., & Tris, K. (2025). Integrating shapley value and least core attribution for robust explainable AI in rent prediction. *Buildings*, *15*(17), 3133. doi: https://doi.org/10.3390/buildings15173133

Zollanvari, A. (2023). Ensemble Learning. In *Machine Learning with Python: Theory and Implementation* (pp. 209-236). Cham: Springer International Publishing.